NoSQL Proof-of-Concept Proposal – Dota2 Game Replay Analysis

Team name: PSG.LGD

Yichun Yan
Ziwei Jiang
Yifan Li
Weiqi Wang

September 29, 2019

Contents

1 Potential Datasets

Our team aims at exploring data about a popular and long-lived computer game, Dota2. We will mainly use two datasets:

1. Dota2 match result dataset
2. Dota2 replay dataset

The first dataset is the snapshot of the final state of each player and the whole game, containing only important information. The second one is a binary format file which can be executed by Dota2's client to reproduce everything happens in the game.

The datasets can be retrieved from the following resources:

1. Valve's Dota2 replay servers
2. Valve's official API

We will get the match results data from Valve's official API. It also provides a key, which we can use to retrieve a replay from the replay server.

In all, we will have two different kinds of data for our further analysis, relatively large-scale but coarse-grained match results data, and very fine-grained replays data. Considering the file size of one replay, we decide to include recent professional games and randomly select some public games for the replay data. WE will mention how we acquire the data in Data ETL section.

2 Five Vs of Datasets

To assess the adequacy of adopting big data to our project, we firstly identify five Vs characteristics of our dataset.

2.1 Volume

Replays are one of datasets we plan to use, which are the full records of finished matches. Size of one replay is estimated to be about 60MB. We plan to collect 3000 replays whose size is about 180GB. As for match result data which is JSON-format data, size of each result is about 5KB. We plan to collect 10,000,000 match results whose size is about 50GB. Moreover, if we desire more data in the future implementation, we can always access it as long as dota 2 still has players.

2.2 Velocity

According to the statistic data in August 2019 from Steam which is a a video game digital distribution platform, the average number of players is 467,148.3 players and the peak number is 826,690. The large number of players makes both the replay data and result data high-velocity. Although, in our project, the velocity depends on the velocity of the API request, this statistic data reveals that it's potential to be high-velocity.

2.3 Variety

From the APIs mentioned above, we can access JSON format files, the match result data, and binary format files, the replay data.

2.4 Veracity

The overall quality of our datasets is good, because we access most of them from the official API. But there is still a few noise. For example, the time of some games is too short, which makes these game not representative. Thus, we plan to implement data validation to filter the undesired matches.

2.5 Value

Our datasets are useful for a relatively long time. Because we plan to collect 1/3 professional matches and 2/3 public matches. The professional matches are durable since we can always extract information about professional teams and players from them. As for the normal games, they are valuable as long as they are not too stale, like more than 2 years ago. In data validation step, we will set a filter on date for normal games which makes our datasets valuable.

3 Potential Business Questions

As mentioned before, we can collect 2 kinds of data, one is match results and one is replays. By analyzing them, the following questions might be answered.

- Easy questions:
  - Who is the hero gaining gold/XP fastest in 15 minutes in normal games/professional games?
  - Who is the hero having most kills/assists/heals/deaths in normal games/professional games?
  - What is the most purchased item in normal games/professional games?
  - Who is the hero having most bad-manner players (players who are AFK or disconnected)?
  - Who is the most popular hero (hero who has highest pick rate) in normal games/professional games?
  - Who is the hero having highest ban rate in professional games?
  - who is the hero having the most ban/pick rate in professional games?
  - How long does a game cost in average in normal games/professional games?
  - Which item is used most in normal games/professional games?
- Moderately challenging questions:
  - How is the benefits gained from buybacks in normal games/professional games
  - How is the vision in normal games/professional games?
  - When does the first team battle happen in normal games/professional games?
  - Which is lineup having highest win rate in normal games/professional games?
  - Which is the most popular lineup (lineup which has the highest pick rate) in normal games/professional games?
  - which is the most common ban-pick combo in professional games?
  - Who is hero changed most in win/pick/ban rate after release new version of Dota2?
- Challenging questions:
  - Does there exist correlation between the time of the first blood and the time of the entire game?
  - Does there exist correlation between the gold/XP source of a hero and its win rate?
  - Does there exist correlation between the distribution of economy and the result of the game?
  - Does there exist correlation between the economic development and the time of the first team battle?

Moreover, we can use answers above to analyze the difference between normal games and professional games as well as the difference between the blind-pick game and the draft-pick game.

4 Data ETL

4.1 Extraction

Valve, the company that develop the game, initially try to provide all the players with easy APIs to access the data of the game. But due to the increasing stress on its data servers, many of its APIs are shut down. At the same time, the documentation seems to be never updated since it's created, which makes it much more difficult for us to collect the data. For example, the API which returns a bunch of the match results given a starting match id is not usable anymore. Another API which returns a key that is critical for construct the URL to download replay of that game is not working as well.

But luckily, we found some hints from the developer's forum of Dota2. We can use the starting sequence number, which works similarly as the match id, to get a series of game results. And by calling a third party API from OpenDota, we can get the important key to construct the URL to download replays again.

The specific steps of gaining our data are as following:

1. Match Results:
    - Use GetMatchHistoryBySequenceNum API to get the match ids. We will need a a field called `start_at_match_seq_num` to specify the starting match sequence number of the results.
    - Then we can extract the last sequence number of the results as the `start_at_match_seq_num` for the next call. By doing this iteratively we can enlarge our dataset for our first kind of data.
2. Replays:
    - Get the match ids of recent professional games from the last step.
    - Use the those ids via OpenDota API to get the information we need (`cluster` and `replay_salt`) for retrieving replays from Valve's replay server
    - Construct links in this format: `http://replay<cluster>.valve.net/570/<match_id>_<replay_salt>.dem.bz2` to get the replays.

4.2 Transformation

The match result is in JSON format so we can almost directly store it. However, the replay data is completely unstructured `.dem.bz2` binary file, data transformation must be done before we can store the data in our database.

Firstly, we can decompress the `.bz2` file with `org.apache.commons.compress.compressors.bzip2` package in Java, which will give us a `.dem` file:

Next, we can utilize clarity, an open source Dota2 replay parser, to extract useful information from the `.dem` file.

It cannot be achieved by a single click, though. We have to write a lot of code to invoke its function. What's more, clarity does not provide a detailed documentation, instead there are only some examples which forces us to iteratively attempt and learn the usage of this tool.

Based on our exploration, the following data will be available:

- Player name, id, team formation and hero choice
- Detailed log of the game, including a hero:
    - deals damage to another one
    - heals another one
    - receives/loses a buff/debuff
    - kills another one
    - uses his ability
    - uses an item
    - buys an item
    - receives/loses some gold
    - gains some XP
    - buys back (spending money in order to instantly re-spawn)
- Spawn/death of heros and NPCs

These information will be organized in a Document and store to our database.

Most fields in our datasets are related to some of the above-mentioned business questions. They are all important. We will generally validate them, for example, fields like the HP, XP, total gold cannot be negative.

What's more important is to filter out invalid data on the game level. Some of the games lasts for only two or three minutes, because some players are disconnected and others just quit the game very quick. We will completely drop these game records.